Diffusion Models

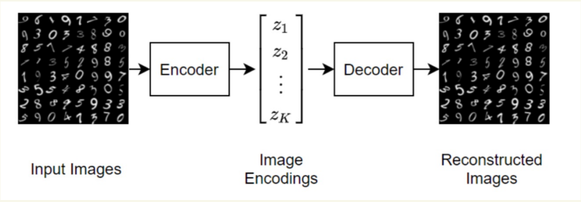
1. Variational Autoencoders
2. Diffusion Models as particular case of VAE
3. SDE
4. LD and FP eq.
5. DM as score-matching
6. Classifier guidance

1 Variational autoencoder (VAE)

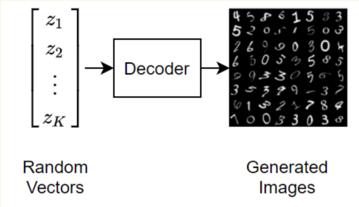
VAE is a neural network that learns to produce its input

Given objects $X = (x_1, .. x_n)$, where $x_i \in \mathbb{R}^D$, construct $z_i \in \mathbb{R}^d$, with which we can reconstruct X.

1.1. Architecture overview



training step
z_i is output of encoder



inference => z_i is a random vector

1.2 Training objective

Training objective : maximize the likelihood of X; $\underline{\underline{\log p(X|\theta) \to \max}}$

For the VAE model : $p(X, z|\theta)$

$$\log p(x|\theta) = \int p(x, z|\theta)\, dz$$

this integral is intractable because relationship between x and z is highly non-linear

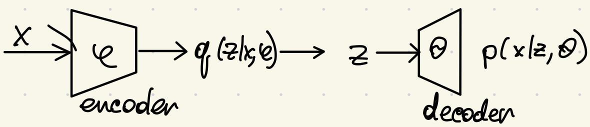
$\int p(x,z|\theta)dz$ is intractable, that's why variational $\mathcal{L}$-bound is used:

$$\log p(x|\theta) \geq \mathcal{L}(q,\theta) = \underbrace{\int q(z|x,\varphi)\log\frac{p(x,z|\theta)}{q(z|x,\varphi)}dz \to \max_{\theta,\varphi}}_{\text{variational inference; ELBO}}$$

$\{\varphi, \theta\}$ = parameters

$$\log p(x|\theta) \geq \underbrace{\int q(z|x,\varphi)\log p(x|z,\theta)dz}_{\text{Reconstruction error}} - \underbrace{\int q(z|x,\varphi)\log\frac{q(z|x,\varphi)}{p(z)}dz}_{\text{KL-divergence (Regularijer)}}$$

1.3 VAE scheme with φ, θ

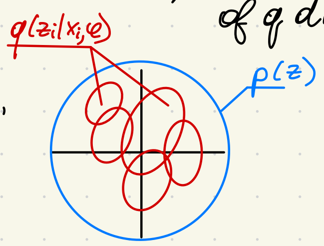


1.4 Meaning of loss components

- Why use regularijer?

To avoid point istributions; with KL term we penalije large deviations of q distributions from z

With KL term, we learn qs close to p(z)!



- Without the reconstruction error, we would have similar qs.

1.5 Disadvantages of VAE

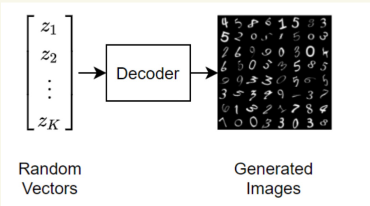
$p(z) \neq \frac{1}{n}\sum_j q_j(z_j|x_j,\varphi)$ → We cannot cover all $p(z)$ with qs.

$p(z) \neq \frac{1}{n}\sum_j q_j(z_j | x_j, \varphi)$ → in generative mode, we take z from prior distribution:

$\hat{z} \sim p(z)$

$x \sim p(x | \hat{z}, \theta)$

(sampling)



That's why in generative mode we do not have realistic output.

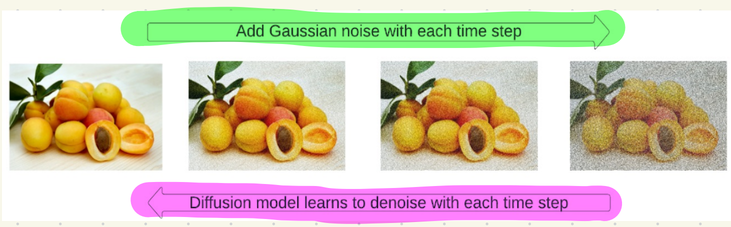
VAEs are good for reconstruction, but not good at generating high-quality images

Exception: if d > D (very deep VAE), we can generate better images

2 Diffusion Models as particular case of VAEs

DMs are also latent variable generative models (like VAEs), but they work by gradually adding noise to the input data:

2.1 Model overview



X_0 X_1 X_{t-1} X_T

Diffusion model is defined by:

- forward
- backward
- sampling

processes

2.2 Diffusion Model modes

Training

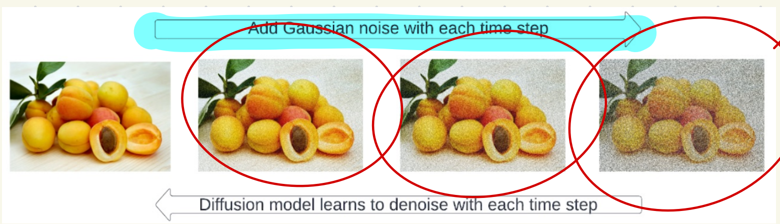
Add noise (forward)
Reconstruct from noisified image (backward)

Inference

Sample from z
Reconstruct from noisy image

2.3 Forward process

latents



$X_0, \dots X_t \rightarrow$ representations of input

At each t step, we add noise

$$X_{t+1} = \sqrt{1-\beta}\, X_t + \sqrt{\beta}\, \varepsilon \qquad \beta \ll 1 \qquad \varepsilon \sim \mathcal{N}(\varepsilon | 0, I)$$

β → constant, ε → noise

in general: $X_t = \sqrt{\alpha_t}\, X_0 + \sqrt{1-\alpha_t}\, \varepsilon$, where $\alpha_t = (1-\beta)^t$

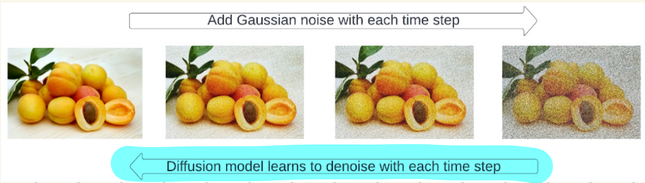
$$\alpha_T \ll 1, \quad \alpha_T \approx 0 \qquad q(X_T | X_0) \approx \mathcal{N}(X_T | 0, I)$$

$$q(X_t | X_0) = \mathcal{N}(X_t | \sqrt{\alpha_t}\, X_0, (1-\alpha_t)) \xrightarrow{t \to T} \mathcal{N}(X_T | 0, I)$$

$$q(x_t|x_0)=\mathcal{N}(x_t|\sqrt{\alpha_t}\,x_0,(1-\alpha_t))\xrightarrow{t\to T}\mathcal{N}(x_T|0,I)$$

Any object from training distribution diffuses to $\mathcal{N}$!
So, DMs do not have VAE's drawback

2.4 Backward process (Training objective



$$\log p(Obs|\theta)\geq\int q(Hid|obs,\varphi)\log\frac{p(Obs,Hid|\theta)}{q(Hid|Obs,\varphi)}$$

VARs $\begin{cases}\text{Observed} = Obs\\ \text{Hidden} = Hid\end{cases}$

Training objective
(see 1.2)

$$\log p(x_0|\theta)\geq\int \underbrace{q(x_1\dots x_T|x_0)}_{\substack{\text{we don't have }\varphi,\\ q \text{ is fixed: } x_t=\sqrt{\alpha_t}x_0+\sqrt{1-\alpha_t}\,\varepsilon}}\log\frac{p(x_0,x_1\dots x_T|\theta)}{q(x_1,\dots x_T|x_0)}dx_1\dots dx_T =$$

$=\}$ apply product rule $\{=$

$$=\int q(x_1\dots x_T|x_0)\log p(x_0|x_1,\theta)dx_1\dots dx_T+$$

$$+\boxed{\int q(x_1\dots x_T|x_0)\log\frac{p(x_1\dots x_T|\theta)}{q(x_1\dots x_T|x_0)}}dx_1\dots dx_T$$

$-KL(q(x_1..x_T)||p(x_1..x_T|\theta))$

$$=\int q(x_1|x_0)\log p(x_0|x_1,\theta)dx_1+$$

$$+\int q(x_1..x_T|x_0)\log\frac{p(x_T)p(x_{T-1}|x_T,\theta)\dots p(x_1|x_2,\theta)}{q(x_T|x_0)q(x_{T-1}|x_T,x_0)\dots q(x_1|x_2,x_0)}dx_1\dots dx_T$$

$$\int q(x_1|x_0)\log p(x_0|x_1,\theta)\,dx_1 + \int q(x_1\dots x_T|x_0)\log\left(\frac{p(x_T)}{q(x_T|x_0)}\frac{p(x_{T-1}|x_T,\theta)}{q(x_{T-1}|x_T,x_0)}\dots\frac{p(x_1|x_2,\theta)}{q(x_1|x_2,x_0)}\right)dx_1\dots dx_T =$$

KLs

$$= \int q(x_1|x_0)\log p(x_0|x_1,\theta)\,dx_1 - \sum_{t=2}^{T}\int q(x_t|x_0)\,KL(q(x_{t-1}|x_t,x_0)\,||\,p(x_{t-1}|x_t,\theta)) - \underbrace{KL(q(x_T|x_0)\,||\,p(x_T))}_{0} \to \max_\theta$$

$0=$

That is why DMs are efficient!

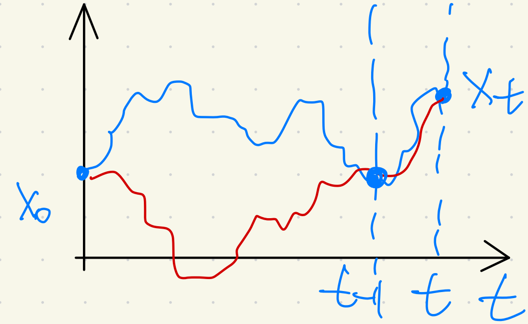
Training objective

$$\log p(x_0|\theta) \geq \mathcal{L}(\theta) \approx \boxed{-\sum_{t=2}^{T}\mathbb{E}_{x_t}KL(q(x_{t-1}|x_t,x_0)\,||\,p(x_{t-1}|x_t,\theta)) \to \max_\theta}$$

$$q(x_{t-1}|x_t,x_0) = \frac{q(x_t|x_{t-1},\cancel{x_0})\,q(x_{t-1}|x_0)}{q(x_t|x_0)} =$$

$$= \frac{q(x_t|x_{t-1})\,q(x_{t-1}|x_0)}{q(x_t|x_0)} =$$

$$= \mathcal{N}(x_{t-1}\,|\,\mu(x_0,x_t),\tilde{\beta}_t I)$$



$$\mu(x_0, x_t) = \frac{\sqrt{\alpha_t}\,\beta}{1-\alpha_t} x_0 + \frac{\sqrt{1-\beta}\,(1-\alpha_{t-1})}{1-\alpha_t} x_t$$

$$\tilde{\beta}_t = \frac{1-\alpha_{t-1}}{1-\alpha_t} \cdot \beta$$

$$q(x_{t-1} \mid x_t, x_0)$$

$$p(x_{t-1} \mid x_t, \theta) = q(x_{t-1} \mid x_t, \underbrace{x_\theta(x_t, t)}_{\text{trained with deep NN}})$$

Then, we can rewrite training objective:

$$KL\big(q(x_{t-1} \mid x_t, x_0) \,\|\, p(x_{t-1} \mid x_t, x_0)\big) = \text{const} \cdot \|x_0 - x_\theta(x_t, t)\|_2$$

2.5 Training procedure

1. Take x_0 from dataset
2. Take arbitrary $\tau \in [2, T]$
3. Generate $x_\tau \sim q(x_\tau \mid x_0)$ $\quad x_\tau = \sqrt{\alpha_\tau}\, x_0 + \sqrt{1-\alpha_\tau}\, \varepsilon$
4. Differentiate $KL\big(q(x_{\tau-1} \mid x_0) \,\|\, p(x_{\tau-1} \mid x_\tau, \theta)\big)$ w.r.t. θ

$$= \text{Const} \cdot \|x_0 - x_\theta(x_\tau, \tau)\|^2$$